Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry Letters 15 (2005) 4076–4084
Bioorganic & Medicinal Chemistry Letters

Modeling K_m values using electrotopological state: Substrates for cytochrome P450 3A4-mediated metabolism

Yong-Hua Wang,[a] Yan Li,[b] Yan-Hong Li,[a] Sheng-Li Yang[a] and Ling Yang[a,*]

[a]*Lab of Pharmaceutical Resource Discovery, Dalian Institute of Chemical Physics, Graduate School of the Chinese Academy of Sciences, #457 Zhongshan Road, Dalian 116023, China*
[b]*School of Chemical Engineering, Dalian University of Technology, #158 Zhongshan Road, Dalian 116012, China*

Received 3 April 2005; revised 26 May 2005; accepted 3 June 2005
Available online 28 June 2005

Abstract—In order to determine K_m values of substrates for CYP3A4-mediated metabolism, an in silico model has been developed in the present work. Using electrotopological state (E-state) indices, together with Bayesian-regularized neural network (BRNN), we have described an in silico method to model log $(1/K_m)$ values of various substrates. The relative importance of the E-state indices is analyzed by principal component analysis. By using an additional external test set, which is independent of the training set, the robustness and predictivity of the model are also validated.

1. Introduction

It is well known that clinical failures of about 30% of the investigational new drug filings are attributed to their inadequate absorption, distribution, metabolism, excretion (ADME), and toxicology attributes. In ADME, the metabolism is known to be a major contributing factor for drug clearance, decreasing free drug concentration and influencing drug pharmacokinetics. Therefore, new chemical entities (NCEs) are often tested early for the metabolism property, to gain details regarding the metabolic fate of xenobiotics in hepatocytes. The cytochrome P450 (CYP450) superfamily is a group of versatile catalysts that play a pivotal role in metabolism of xenobiotic and endogenous compounds. Of major importance is the involvement in drug metabolism and toxicant biotransformation[1]. Thus, intense efforts have been focused on the study of CYP isoform(s)-mediated metabolism, and hence prediction of degrees likely to alter elimination of drugs in vivo[2]. Accordingly, obtaining information (e.g., K_m and V_{max}) of an NCE for CYP isoform(s)-mediated metabolism is of significance.

As a major isozyme of P450, CYP3A4 constitutes 40–60% of the human hepatic P450 complement[3], metabolizing 40–50% drugs in clinical use[4,5]. Therefore, CYP3A4 has become an important subject of study on drug metabolism. At the present time the majority of metabolism investigations are carried out in vitro by using hepatic cells or recombinant CYP3A4 enzymes. The data derived from in vitro models are valuable because they can be extrapolated to humans in vivo, although it is still a complex process. Moreover, informative parameters such as K_m, are also useful as they relate to optimizing the drug-like properties of promising drug leads. However, extensive experimental studies to obtain this information are expensive and require collaborations between different groups[4].

As a facile and economic alternative to in vitro methods, in silico modeling techniques are now emerging. In this paper, we propose a computational model for the enzyme kinetics of CYP3A4. Molecular electrotopological state (E-state) descriptors serve as the basis for structure representation of molecules in the present case. E-state indices have been applied in building various QSAR models.[5,6] Our experience clearly indicates their usefulness in QSAR modeling. Moreover, the newly developed hydrogen E-state indices[7] have facilitated the capability of the E-state descriptors. Bayesian-regularized neural network (BRNN)[8,9] was employed as a regression tool to generate the QSAR models for K_m, with a data set

Keywords: CYP3A4; log $(1/K_m)$; Electrotopological state indices; Bayesian-regularized neural network.
* Corresponding author. Tel.: +86 0411 84379317; fax: +86 0411 84676961; e-mail: yling@dicp.ac.cn

0960-894X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2005.06.015

of 59 various CYP3A4 substrates [10]. The validity of the computational model was also demonstrated by use of an additional test that is independent of the training one.

2. Results and discussion

2.1. Molecular descriptors

The original 50 components have been compressed and analyzed by the principal component analysis (PCA), resulting in 14 principal components (PCs). Table 1 depicts the statistical results of the PCA. In the present paper, the number of PCs was determined by the standard error of prediction of the test set rather than by the maximum variance described by the PCs. The 14 PCs obtained were sufficient to explain 95.03% of variance and were sufficient in the network-modeling cases reported here. The coefficients for these 14 PCs are shown in Table 1. The first PC (PC1) explains 38.15% of the total variance, and each component is expected to contribute in an equal manner to PC1. The second PC, accounting for about 13.97% of the total variance, consists mainly of the nwHBa [the number of weak hydrogen-bond (H-bond) acceptors]. The third PC, explaining about 8.14% of the total physical properties of descriptors, is dominated by the number of atom type $\gtrdot C \lessdot$ descriptor. PC4, answering for 7.25% of the total variance, can be attributed to E-state values in descending order as follows: methylenes, N atom group, the largest hydrogen E-state value, and the descriptors of molecular internal hydrogen bonds. PC5, which is responsible for 5.57% of the total variance, is dominated by the factors describing extreme atom level E-state values in the molecule, which is mainly contributed from the largest hydrogen E-state value. The remaining nine PCs, accounting for 21.95% of the variance, have strong contributions from 7 to 25 different E-state descriptors.

Though cytochrome P450 3A4 metabolizes a wide variety of compounds, the protein exhibits certain substrate specificity.[11] Recent studies have indicated that steric constraints exerted by the enzyme as well as the inherent chemical reactivity of the substrate are important determinants of regioselectivity of hydroxylations catalyzed by P450 3A4.[12] Clearly, the specific interaction determines the selectivity of an enzyme to its ligand, which is often caused by noncovalent binding. Therefore, for a successful topologically based QSAR model, the structural descriptors should be responsible for the noncovalent interactions between the ligand and the receptor. Interestingly, essential features of noncovalent interactions have been encoded in E-state indices.[13] In E-state representation, the intrinsic state term, I_i, encodes in an integrated manner both topological (also called steric attributes) and electronic attributes. The intrinsic state is derived from the ratio of valence state electronegativity of an atom to a measure of its local topological character. The E-state indices describe the electron accessibility at each atom, whose attributes are very significant in noncovalent interactions crucial to the development of QSAR models. Thus, in E-state representation, the attributes of structure, essential in describing the molecular interactions and proving the basis for relationships, are integrated in QSAR models based on the E-state.[13]

2.2. QSAR modeling

The QSAR modeling results based on the BRNN for log $(1/K_m)$ are summarized in Table 2 (training) and Table 3 (test). The calculated (calcd) log $(1/K_m)$ values and residuals (res) are also given in Tables 2 and 3. Figure 1 shows the performance of the BRNN model for the training and test data, which indicated that the samples are well distributed around the line. For the training set the standard error of estimation (SEE) is 0.11 ± 0.02 (data scaled from 0 to 1) and for the testing set the standard error of prediction (SEP) is 0.15 ± 0.02 (two outliers omitted). Both SEP of the test data and SEE of the training data are of the same order of the magnitude, indicating that the model is not overfitted. In addition, the E-state indices correlated well with the log $(1/K_m)$ values ($R^2 = 0.66 \pm 0.03$ and $R^2 = 0.40 \pm 0.04$ for the training and test data, respectively). A study of the residues of training and testing data (Tables 2 and 3) revealed some compounds whose difference between calculated and actual log $(1/K_m)$ values was larger than one logarithmic unit. The most obvious outliers were felodipine and zileuton, for both of which the prediction error was greater than one logarithmic unit in the training models. However, the two compounds were still included in training the model. In test models, two compounds, namely senecionine and venlafaxine, seem to be outliers, as the residues are considerably large (shown in Table 3). The structures of the two compounds are shown in Figure 2. The exclusion of the outliers in the test set improved the correlations significantly. The possible reason for the two compounds being outliers is that the original observed data might not be very accurate. Sun has addressed quality of data for ADMET modeling in his recent review.[14] Biological data often generate uncertainty and tend to have large variation. Accuracy refers not only to experimental errors, but also to uniformity of experimental conditions, especially when data must be collected from different sources.[14] From the results, it can be concluded that the present method, that is, the BRNN, is predictive from both the internal (training) and external point of views with respect to the predictions of the test sets.

2.3. Model interpretation

In order to validate the robustness of the BRNN model, different numbers of hidden neurons from 10 to 25 were tried and no significant deviation of SEP was found for each prediction (the standard deviation <0.05; data not shown). The results illustrate one of the important advantages of the Bayesian neural nets. The models obtained were relatively independent of the architecture of the nets, provided that a minimum number of hidden layer nodes were used. The resulting optimum BRNN model for the present data has a 14–13–1 architecture.

From the examination of the plot (Fig. 1), we can find that the calculated errors of the test sets are of the same

Table 1. Descriptive statistics for the 14 principal components[a]

Descriptor	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	PC 11	PC 12	PC 13	PC 14
nHBd	0.044	0.030	0.058	−0.047	0.020	0.113	−0.008	−0.120	−0.065	0.148	−0.055	0.021	−0.028	−0.037
nHBa	0.047	0.038	0.002	−0.028	−0.058	−0.019	0.067	0.070	−0.004	−0.025	0.061	0.051	0.101	0.151
nwHBa	0.002	**0.121**	0.043	0.038	0.088	−0.110	0.099	0.085	0.012	−0.009	−0.009	0.117	0.050	0.082
SHBd	0.046	0.020	0.074	−0.030	0.015	0.088	−0.018	−0.111	−0.061	0.116	−0.050	0.064	−0.031	0.005
SHBa	0.050	0.025	0.016	−0.020	−0.029	−0.039	0.018	0.059	−0.036	−0.064	0.050	−0.013	0.055	0.025
SwHBa	−0.023	0.089	0.056	0.093	0.088	−0.093	−0.010	0.009	0.144	0.208	−0.001	0.206	−0.030	−0.063
H_{max}	0.039	−0.011	0.061	−0.005	0.120	0.144	0.018	−0.033	0.201	−0.064	−0.003	0.080	−0.019	−0.035
G_{max}	0.034	0.005	0.038	−0.033	0.079	0.088	0.096	0.172	0.000	−0.219	0.073	**−0.419**	−0.018	−0.187
Hmin	−0.004	−0.011	0.066	−0.130	0.062	−0.031	−0.094	**0.469**	0.254	0.179	0.130	0.305	0.167	−0.191
Gmin	−0.036	−0.008	−0.057	0.030	−0.065	0.009	−0.037	−0.165	0.180	0.095	−0.296	0.189	0.051	0.567
$H_{max\ pos}$	0.039	−0.011	0.061	−0.005	0.120	0.144	0.018	−0.033	0.201	−0.064	−0.003	0.080	−0.019	−0.035
SHBint2	0.005	0.054	0.066	−0.126	0.115	0.049	−0.047	0.073	0.077	0.181	0.123	−0.137	0.258	**0.588**
nHaaCH	−0.019	0.115	0.084	0.042	0.012	−0.091	−0.053	−0.089	0.010	0.023	0.045	−0.108	−0.037	−0.075
nHCsats	0.045	−0.007	−0.033	0.099	−0.076	0.026	−0.064	−0.007	0.039	0.110	0.029	−0.155	0.012	0.036
nHCsatu	0.040	0.021	−0.066	0.015	0.039	−0.102	0.046	−0.153	0.182	−0.121	−0.073	0.212	0.119	−0.332
$nsCH_3$	0.043	0.010	−0.060	0.015	−0.137	−0.025	−0.026	0.044	0.037	0.061	−0.170	0.029	−0.187	−0.031
$nssCH_2$	0.025	0.005	−0.037	0.191	0.092	0.081	−0.027	−0.059	0.018	−0.022	0.125	−0.163	0.330	−0.144
ndsCH	0.029	−0.024	−0.094	0.042	0.193	−0.116	0.078	0.044	−0.129	0.180	0.058	0.025	−0.194	0.115
naaCH	−0.019	0.115	0.084	0.042	0.012	−0.091	−0.053	−0.089	0.010	0.023	0.045	−0.108	−0.037	−0.075
nsssCH	**0.048**	−0.014	−0.019	0.024	−0.066	−0.058	−0.048	−0.105	0.124	0.091	0.071	−0.038	0.015	−0.006
ndssC	0.041	0.025	−0.051	−0.021	0.067	−0.086	0.058	0.002	−0.028	−0.241	−0.194	0.150	0.237	0.178
naasC	−0.015	0.057	0.033	0.035	−0.060	0.088	0.340	0.222	−0.186	−0.070	−0.092	0.073	0.074	−0.072
nssssC	0.023	−0.020	**0.131**	0.035	0.058	0.139	−0.134	0.222	−0.138	0.058	−0.189	0.183	0.034	−0.041
nssNH	0.018	0.093	−0.078	−0.109	−0.005	0.141	0.065	−0.123	−0.051	0.149	−0.080	−0.032	0.071	−0.010
naaN	−0.008	0.020	−0.027	0.007	−0.022	**0.160**	0.144	−0.184	0.155	−0.196	**0.663**	0.342	−0.352	0.168
nsssN	0.015	0.068	−0.106	0.107	−0.083	0.064	−0.154	0.238	0.022	−0.067	0.047	0.174	0.005	0.155
nsOH	0.042	−0.034	0.114	0.037	0.016	0.030	−0.044	−0.111	−0.013	0.039	−0.037	0.125	−0.079	−0.021
ndO	0.042	0.049	−0.068	−0.063	0.024	−0.047	−0.061	0.099	0.001	−0.147	0.026	−0.076	0.018	−0.156
nssO	0.028	−0.036	0.067	0.040	−0.145	−0.079	**0.234**	0.066	0.095	0.236	0.168	−0.124	0.157	0.049
SHsOH	0.042	−0.031	0.118	0.032	0.012	0.024	−0.041	−0.099	−0.019	0.032	−0.030	0.131	−0.073	0.009
SHssNH	0.019	0.094	−0.078	−0.106	−0.004	0.132	0.054	−0.108	−0.056	0.142	−0.079	−0.024	0.072	−0.003
SHdsCH	0.029	−0.021	−0.089	0.039	**0.194**	−0.114	0.092	0.058	−0.154	0.194	0.056	0.019	−0.181	0.133
SHaaCH	−0.018	0.116	0.090	0.039	0.011	−0.085	−0.038	−0.069	−0.002	−0.002	0.059	−0.103	−0.017	−0.066
SHCHnX	−0.007	−0.031	−0.031	−0.060	−0.066	−0.053	−0.136	−0.160	**−0.444**	0.129	0.358	0.334	**0.479**	−0.243
SHCsats	0.047	−0.004	−0.003	0.080	−0.088	0.003	−0.041	0.020	0.015	0.118	0.045	−0.087	0.019	0.061
SHCsatu	0.043	0.030	−0.053	−0.004	0.009	−0.109	0.022	−0.123	0.174	−0.135	−0.031	0.181	0.140	−0.253
$SsCH_3$	0.042	0.008	−0.067	0.031	−0.133	−0.013	−0.029	0.019	0.045	0.083	−0.207	0.015	−0.189	−0.065
$SssCH_2$	−0.001	−0.007	−0.048	**0.212**	0.079	0.147	−0.049	−0.060	0.037	−0.031	0.083	−0.232	0.333	0.007
SdsCH	0.029	−0.026	−0.096	0.058	0.187	−0.114	0.070	0.030	−0.117	0.174	0.053	0.057	−0.177	0.097
SaaCH	−0.021	0.111	0.074	0.063	0.007	−0.092	−0.062	−0.107	0.019	0.069	0.005	−0.052	−0.064	−0.054
SsssCH	−0.044	−0.015	−0.036	0.033	0.129	0.101	0.044	0.021	−0.041	−0.013	−0.025	−0.073	0.050	−0.037
SdssC	−0.025	−0.054	0.019	0.083	0.085	0.022	0.148	−0.122	0.075	0.051	−0.254	0.212	0.310	−0.064
SaasC	−0.025	0.056	0.004	0.098	−0.067	0.092	0.220	0.104	−0.100	−0.001	−0.144	0.262	−0.078	−0.249
SssssC	−0.032	−0.010	−0.117	−0.054	0.007	−0.008	0.028	−0.003	0.388	0.155	0.040	−0.032	0.196	−0.135
SssNH	0.017	0.093	−0.074	−0.101	0.000	0.152	0.076	−0.113	−0.074	0.169	−0.059	−0.022	0.067	0.036
SsssN	0.007	0.068	−0.069	0.157	−0.070	0.095	−0.129	0.237	−0.042	−0.051	0.093	0.173	0.047	0.300
SsOH	0.043	−0.028	0.116	0.041	0.008	0.028	−0.043	−0.103	−0.040	0.034	−0.033	0.127	−0.075	−0.004
SdO	0.043	0.050	−0.070	−0.054	0.012	−0.049	−0.066	0.099	0.001	−0.135	0.031	−0.053	0.017	−0.145
SssO	0.029	−0.036	0.070	0.043	−0.149	−0.074	0.220	0.063	0.088	0.239	0.150	−0.118	0.156	0.048
SaOm	0.027	−0.004	0.079	−0.029	−0.012	−0.150	0.095	−0.057	−0.069	**−0.398**	−0.033	0.045	0.236	0.500
Eigenvalue	19.08	6.99	4.07	3.63	2.79	2.38	1.95	1.21	1.19	1.07	0.92	0.84	0.80	0.60
%VE[b]	38.15	13.97	8.14	7.25	5.57	4.77	3.91	2.42	2.40	2.13	1.85	1.67	1.60	1.20
TVE[c]	38.15	52.12	60.25	67.51	73.08	77.85	81.76	84.18	86.58	88.71	90.55	92.23	93.83	95.03

[a] The most important descriptors in each PC are shown in bold.
[b] % of Variance explained.
[c] Total % variance explained.

order of magnitude as experimentally observed, ranging from 0.1 to 32.5. Although some errors are relatively large, these values could be expected to relate to the reasonable experimental values for those compounds. Further, the SEP for the test set and the SEE for the training set are of the same order of magnitude. These results indicate another advantage of the BRNN. The BRNN-based model shows robustness in its predictions, as the network allows evaluation of the likely uncertainty of a prediction.[15] The Bayesian framework deals with uncertainty by applying probabilities to each possible event. Therefore the objective function parameters of the net can be estimated using statistical techniques.[15] Moreover, BRNN applies a posterior distribution over network weights that will give rise to a distribution over the outputs of the network for some new case, so that a

Table 2. CYP3A4 substrates and their log $(1/K_m)$ values of the training set[a]

No.	Name	Pathway	Observed[b]	Calculated[c]	Residual[d]	References
1	Clarithromycin	14-(*R*)-Hydroxylation	−1.6875	−1.239	0.4485	10a
2	Adinazolam	Hydroxylation	−1.9628	−2.029	−0.0662	10b
3	Carbamazepine	10,11-Epoxidation	−2.6532	−2.0185	0.6347	10c
4	Colchicine	3-Demethylations	−3.7482	−2.7903	0.9579	10d
5	Dexamethasone	6β-Hydroxylation	−1.3655	−1.5889	−0.2234	10e
6	Felodipine	Ring oxidation	−0.8388	−1.889	**−1.0502**[e]	10f
7	Indinavir	All metabolites	−0.1139	−0.5216	−0.4076	10g
8	Mifepristone	Oxidation	−1.0212	−0.9257	0.0955	10h
9	Onapristone	N-Demethylation	−1.2923	−0.6936	0.5986	10i
10	Rapamycin	41-O-Demethylation	−0.9069	−1.2256	−0.3187	10j
11	Saquinavir	M2+M7	0.4559	−0.3118	−0.7677	10k
12	Tamoxifen	N-Demethylation	−1.9912	−2.0561	−0.0648	10l
13	Tazofelone	Oxidation	−0.9955	−1.5751	−0.5797	10m
14	Flunitrazepam	N-Demethylation	−2.1892	−1.5864	0.6028	10n
15	Zileuton	Sulfoxidation	−2.9138	−1.86	**1.0538**[e]	10o
16	Bromodichloromethane	—	−1.4472	−1.5719	−0.1247	10p
17	Amitriptyline	Hydroxylation	−1.8407	−2.1605	−0.3198	10q
18	Citalopram	Demethylation	−2.4548	−2.3039	0.151	10r
19	Cyclosporine	All metabolites	−0.6021	−0.3834	0.2187	10s
20	Ethoxycoumarin	O-Deethylation	−3.4771	−2.8261	0.651	10t
21	Haloperidol	All metabolites	−1.3874	−1.3812	0.0062	10u
22	K 02	Hydroxylation	−1.4843	−1.3782	0.1061	10v
23	1-Nitropyrene	Tatal metabolism	−0.5185	−1.2443	−0.7258	10w
24	Pimozide	N-Dealkylation	−0.7559	−1.5804	−0.8245	10x
25	Ritonavir	M1+M2+M3	−1.301	−1.0659	0.2351	10y
26	Simvastatin	Hydroxylation	−1.3201	−1.7046	−0.3845	10z
27	Taxol	6α-Hydroxylation	−1.2553	−0.8004	0.4548	10aa
28	Tirilazad mesylate	Hydroxylation	−0.2601	−1.0799	−0.8199	10ab
29	Vindesine	Tatal metabolism	−1.3927	−1.274	0.1187	10ac
30	Etoposide	—	−1.5366	−1.5056	0.0309	10ad
31	Amiodarone	N-Demethylation	−2.4914	−2.1744	0.3169	10ae
32	Bupivacaine	N-Dealkylated	−2.0967	−1.5058	0.5909	10af
33	Clozapine	N-Oxidation	−1.7853	−1.9663	−0.181	10ag1,10ag2
34	Dapsone	N-Hydroxylation	−2.2304	−2.1476	0.0829	10ah
35	Ezlopitant alkene	Oxidation	−1.3424	−1.5056	−0.1632	10ai
36	Imipramine	N-Demethylation	−2.0398	−1.7652	0.2746	10aj
37	Lilopristone	N-Demethylation	−0.9445	−1.1345	−0.1901	10i
38	Omeprazole	Hydroxylation	−1.6902	−2.4943	−0.8041	10ak
39	Quinine	3-Hydroxylation	−2.0257	−1.7071	0.3186	10al
40	*S*-Berapamil	Norverapamil+D-617	−1.6902	−2.3982	−0.708	10am
41	Tacrolimus	31-O-Demethylation	−0.7924	−0.88	−0.0876	10an
42	Taxotere	All metabolites	−0.0414	−0.7103	−0.6689	10ao
43	Troglitazone1	—	−2.7918	−1.8457	0.9461	10ap
44	Zatosetron	Oxidation	−2.3901	−1.8413	0.5487	10aq

[a] Mean of multiple measurements from the same or different data. — Not given in the literature.
[b] Observed log $(1/K_m)$ value.
[c] log $(1/K_m)$ value calculated by QSAR model.
[d] log $(1/K_m)$−calculated.
[e] Outliers are shown in bold.

local minimum is likely to be avoided. Therefore, compared with traditional statistics (e.g., standard least squares), the Bayesian neural networks are able to find the more general relationships between structural and activity data.

Several studies have demonstrated that in BRNN modeling the SEE is sufficient to estimate the performance of the network's generalization.[16,17] One more advantage of BRNN is that although no validation or test is involved the models are still robust and make optimal predictions.[18–20] However, for a small data set, we believe an additional test set may be helpful to validate the network behavior. For the present data, we used an external test set containing one-fourth of the whole data to evaluate the model. As can be seen from Figure 1, the model shows a reasonable prediction for the external data set (Table 2), although no validation set was involved in model building. The indications are that BRNN-based model appears reasonable and is likely to give useful estimation of log $(1/K_m)$ values for new data.

3. Conclusion

For CYP3A4 mediated-metabolism of substrates, a successful BRNN-based QSAR model was developed with

Table 3. CYP3A4 substrates and their log ($1/K_m$) values of the test set[a]

No.	Name	Pathway	Observed[b]	Calculated[c]	Residuals[d]	References
1	Erythromycin	N-Demethylation	−1.7853	−1.0079	−0.7774	24a
2	Alprazolam	4- Hydroxylation	−2.4232	−2.1431	−0.2801	24b
3	Carteolol	Oxidation	−1.8432	−1.4867	−0.3565	24c
4	Cortisol	6β- Hydroxylation	−1.1818	−0.8282	−0.3536	24d
5	Diltiazem	N-Demethylation	−1.3617	−1.1168	−0.2449	24e
6	Fentanyl	Norfentanyl formation	−2.0682	−1.5403	−0.5279	24f
7	Irinotecan	Hydroxylation	−1.2648	−1.0585	−0.2063	24g
8	Vinblastine	Metabolite M formation	−0.8338	−1.4946	0.6608	10ac
9	Pimobendan	O-Demethylation	−2.4024	−1.9873	−0.4151	24h
10	Rifabutin	All metabolites	−1.0373	−1.8754	0.8381	24i
11	Senecionine	N-Oxidation	−2.4472	−1.4166	**−1.0306**	24j
12	Taurochenodeoxycholic acid	Hydroxylation	−1.9542	−2.0561	0.1019	24k
13	Terfenadine	Parent consumption	−0.9823	−1.1716	0.1893	24l
14	Venlafaxine	N-Demethylation	0.0899	−1.458	**1.5479**	24m
15	Chlorzoxazone	6-Hydroxylation	−2.2040	−2.1766	−0.0274	10t

[a] Outliers are shown in bold.
[b] Observed log ($1/K_m$) value.
[c] log ($1/K_m$) value calculated by QSAR model.
[d] log ($1/K_m$)−calculated.

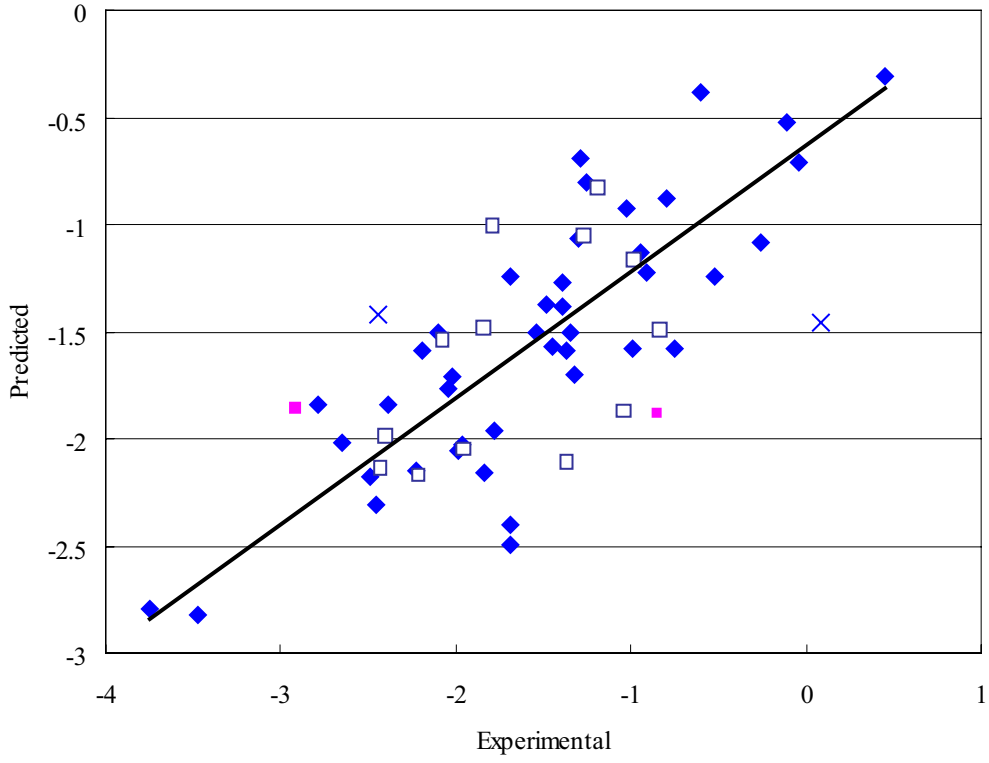


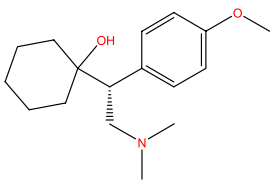
Figure 1. Plot of calculated log ($1/K_m$) values (QSAR model) versus observed log ($1/K_m$) for the substrates for CYP3A4-mediated metabolism. (♦) training data; (□) testing data; (■) outlier in training set; (×) outlier in test set.

E-state structure descriptors. These descriptors include 50 theoretical parameters, encoding important structural characteristics that are useful for analyzing the enzyme–ligand interaction. By this QSAR model, the numerical E-state descriptors can be converted to log ($1/K_m$) values of new compounds, and thus the binding affinity of a ligand to enzyme can be easily obtained. Since the robust model employs large diversity of molecules for the training set and can avoid the difficulty of measuring the ADME properties, it should be of use in further drug evaluation and design.

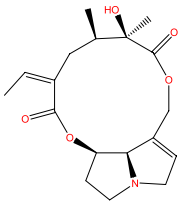
4. Methods

4.1. Data set

Data of compounds involved in CYP3A4-mediated metabolism have been measured experimentally by several investigators. For the present investigation, we limited the data to in vitro measurements with human liver microsome or cDNA-expressed human CYP3A4[10] (Table 1). Although the experimental conditions are not identical, we consider them sufficiently similar to be



venlafaxine



senecionine

Figure 2. The chemical structures of venlafaxine and senecionine.

included in the same data set. If a successful model is obtained, the inclusion of all the data appears justified. The resulting data set consists of 59 observations on neutral compounds with their corresponding CYP3A4 enzyme kinetic parameters (K_m), which cover a wide variety of CYP3A4 substrates. These K_m values were converted to logarithm basis for analysis. See Tables 2 and 3 for the experimentally derived kinetic parameters. The log ($1/K_m$) values cover a very large range, with over 4 orders of magnitude from −3.75 to +0.46.

In the present work, using a *K*-means clustering algorithm on the *X* (indices) and *Y* (log ($1/K_m$)) values, the data set was divided into two sets, one for training and one for testing. The maximum and the minimum log ($1/K_m$) values were put in the training set, considering that the training set should cover the whole range of log ($1/K_m$) values to avoid extrapolation during the final prediction. One-fourth of the total number of compounds were arranged in the test set, which were not used in training the regression models. The remaining compounds were used as the training set.

4.2. E-state indices generation

For the present case, molecular structure descriptors known as electrotopological state (E-state) indices were employed. These indices have been applied in many QSAR studies where the E-state formalism has been described.[5–7] The nature of the E-state indices is such that they encode not only the information about the topological environment of an atom, but also the electronic interactions from other atoms in the molecule. Thus, E-state is able to provide useful information on structure features that mostly relate to the property to be modeled[5–7]. The E-state indices have been used in both their atom-level and atom-type forms. In the present work, the above forms of E-state indices were applied to represent the molecules (Table 4), which also include hydrogen E-state indices, HES_i, a more recent extension of the E-state. Two-dimensional structures of all substrates under study were obtained from a commercial available MDL-ISIS database. The E-state indices were calculated using the Molconn-Z program in SYBYL soft package (Tripos Associates. St. Louis, MO).

4.3. Feature reduction

To eliminate noise (uninformative descriptors) and prevent overfitting or chance correlations in building models, we reduce the number of descriptors by the following procedures: (i) Descriptors with constant values as well as descriptors containing 95% zero values were removed, because they offered little information for the construction of the models. Following this step, 50 descriptors were obtained as shown in Table 1. (ii) To further reduce the chance of correlation among descriptors and the descriptor space, a principal component analysis was performed on the 50 components (descriptors). Before the PCA process, the given data set (including structural feature data and activity data) was normalized so that the input and output variables would have means of zero and standard deviations of one. After this data matrix reduction procedure, 14 principal component vectors were constructed and used further for statistical analysis in model development. The main calculations of PCA were done by a free software Tanagra 1.1 offered by Rakotomalala (http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html).

4.4. Bayesian regularized neural network

The data matrix was submitted for QSAR modeling using the BRNN, a robust SAR mapping method. Although the Bayesian concept and methodology have existed for many years, they have gained popularity as a tool for facilitating drug development only in recent years.[8,9] Since the Bayesian method has been reported in papers by Mackay[18,19,21] and Buntine and Weigend,[20] only a brief summary of the method is presented here. BRNNs are multiplayer feed-forward neural networks trained with Bayesian algorithm.

Such a net uses a set of network weights and considers a probability distribution of net weights.[18,19,21] In contrast to conventional networks (e.g., classical back-propagation neural networks, BPNN) using a single set of weights to make predictions of new input data, the Bayesian training produces a posterior distribution over network weights. By Bayesian inference the most plausible set of weights that match the data well occurs in the maximum posterior distribution.[18,19,21]

To make it clearer, we present a brief process of the network. Firstly, before the data are observed, the true relationship is expressed in a prior probability distribution over the network weights that define this relationship. After observation of the data, a posterior distribution $P(W|D,H_i)$ over network weights is determined by Bayesian inference, and the related properties from the prior probability distribution $P(W|H_i)$ provided by the training set *D* using the model *H* are also determined. The probability distribution is made by a Gaussian approximation. If a Gaussian prior distribution that penalizes net weights is applied, and the data are assumed by a smooth function with additive Gaussian

Table 4. Electrotopological state indices used in this work

Variable	Definition	Variable	Definition
NHBd, nHBa, nwHBa	nHBa, nHBd, nwHBa Hydrogen bond donor and acceptor counts (nwHBa are the weak hydrogen bonds)	ShsOH, SHssNH, SHdsCH SHaaCH SHCHnX, SdsCH SaaCH, SsssCH	Atom type electrotopological state index values for atom types
H_{max}, G_{max}	Extreme atom level E-state values in molecule:	SHCsats	E-state of Csp3 bonded to other
H_{min}, G_{min}	H_{max}—Largest hydrogen E-state value	SHCsatu	Saturated C atoms. E-state of Csp3
$H_{max\,pos}$	G_{max}—Largest E-state value		Bonded to unsaturated C atoms
	H_{min}—Smallest hydrogen E-state value		
	G_{min}—Smallest E-state value		
SHBint2	Descriptors of potential internal hydrogen bonds. Internal hydrogen bond descriptor is the product of hydrogen E-state value and E-state value.	SHBa	Acceptor descriptor for molecule (sum of E-state values for all hydrogen bond acceptors in the molecule). The following groups are classified as acceptors: –OH, =NH, –NH2, –NH–, >N–, –O–, =O, –S– along with –F, and –Cl.
SHBd	Donor descriptor for molecule (sum of hydrogen E-state values for all hydrogen bond donors in the molecule). The following groups are classified as donors: –OH, =NH, –NH2, –NH–, –SH, and #CH.	SwHBa	Descriptor for weak hydrogen bond acceptor (sum of E-state values for all weak hydrogen bond acceptors).
SsOH, SdO, SssO, SaOm	Atom type electrotopological state index values for atom types	SdssC, SaasC, SssssC, SssNH, SsssN	Atom type electrotopological state index values for atom types
$nsCH_3$	Number of group atom $–CH_3$	nHCsats	CHn (saturated)
		nHCsatu	CHn (unsaturated)
$nssCH_2$	$–CH_2–$	nssNH	–NH–
ndsCH	=CH–	naaN	:N:
naaCH	:CH:	nsssN	>N–
nsssCH	>CH–	nsOH	–OH
ndssC	=C<	ndO	=O
naasC	:C:–	nssO	–O–
nssssC	>C<	nHaaCH	Number of :CH:
$SsCH_3$	Methyls	$SssCH_2$	Methylenes

noise, maximizing the posterior distribution is equal to minimizing the standard sum-of-squares error together with a weight decay regularizer[22]. An input vector combined with the posterior weight distribution generates a distribution of network outputs. The mean μ and the variance σ^2 of a Gaussian approximation to this predictive distribution can then be calculated to provide $\pm\mu$ on the mean prediction μ.

In this work, this Bayesian regularization takes place within the Levenberg–Marquardt algorithm (LMBR). The combination of the two methods can accelerate convergence of targets and determine the optimum weights for the network.[19,20,23] A brief introduction of the LMBR is also presented here. For a feed-forward neural network with Bayesian regularization, the performance function, which is commonly used is

$$F = \beta F_e + \alpha F_w,$$

where F_w is the mean sum of squares of the network weights, α and β are objective function parameters,[23] and F_e can be expressed as

$$F_e = \frac{1}{n}\sum_{i=1}^{n}(t_i - a_i)^2,$$

where t_i and a_i are the targets and the networks outputs, and n is the number of observations. As different from the conventional feed-forward networks, an additional term F_w is added in the performance function to improve the network generalization. The Bayesian regularization proposed by Mackay[18] automatically sets optimum values for objective function parameters. The Levenberg–Marquardt optimization algorithm, as a Newton step-based method, exhibits quadratical speed of convergence. When training networks with the LMBR algorithm, the combination of squared errors and weights, F, was minimized; therefore, the network weight and bias can be updated. In this case, the Bayesian regularization can ensure correct combination to produce a network that generalizes well.

In the present work, we adopted a three-layer fully connected BRNN with 13 neurons in the hidden layer and one in the output layer. A tan-sigmoid function and a linear transfer function were used for the hidden and output neurons, respectively. The training is stopped at the maximum of the evidence maximum for the hyperparameters α and β (α represents the weight decay regularization, while β governs the variance of the noise).[17] Using BRNN, with an optimal number of

PCs and architecture, the neural network models were trained independently 25 times to eliminate spurious effects caused by the random set of initial weights. We used an internally developed C language program based on the paper by Foresee and Hagan.[23]

Acknowledgments

The authors thank the 973 Program (2003CCA03400) and the 863 Program (2003AA223061) of the Ministry of the Science and Technology of China, and DICP Innovation Fund of Chinese Academy of Sciences.

References and notes

1. Wrighton, S. A.; Stevens, J. C. *Crit. Rev. Toxicol.* **1992**, *22*, 1.
2. Smith, D. A. High throughput drug metabolism. In *Drug Metabolism Toward the Next Millennium*; Gooderham, N. J., Ed.; IOS press: Amsterdam, 1998, pp 137–143.
3. Shimada, T.; Yamazaki, H.; Mimura, M.; Inui, Y.; Guengerich, F. P. *J. Pharmacol. Exp. Ther.* **1994**, *270*, 414.
4. Clish, C. B.; Davidov, E.; Oresic, M.; Plasterer, T. N.; Lavine, G.; Londo, T.; Meys, M.; Snell, P.; Stochaj, W.; Adourian, A.; Zhang, X.; Morel, N.; Neumann, E.; Verheij, E.; Vogels, J. T.; Havekes, L. M.; Afeyan, N.; Regnier, F.; van der Greef, J.; Naylor, S. *OMICS* **2004**, *8*, 3.
5. Wang, Y.; Han, K.; Yang, S.; Yang, L. *J. Mol. Struc. (THEOCHEM)* **2004**, *710*, 215.
6. Huuskonen, J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 425.
7. Rose, K.; Hall, L. H. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 651.
8. Winkler, D. A.; Burden Frank, R. *J. Mol. Graph. Model.* **2004**, *22*, 499.
9. Wang, Y.; Li, Y.; Yang, S.; Yang, L. *J. Comput. Aid. Mol. Des.*, in press.
10. (a) Rodrigues, A. D.; Roberts, E. M.; Mulford, D. J.; Yao, Y.; Ouellet, D. *Drug Metab. Dispos.* **1997**, *25*, 623; (b) Venkatkrishnan, K.; von Moltke, L. L.; Duan, S. X.; Fleishaker, J. C.; Shader, R. I.; Greenblatt, D. J. *J. Pharm. Pharmacol.* **1998**, *50*, 265; (c) Kerr, B. M.; Thummel, K. E.; Wurden, C. J.; Klein, S. M.; Kroetz, D. L.; Gonzalez, F. J.; Levy, R. H. *Biochem. Pharmacol.* **1969**, 47; (d) Tateishi, T.; Soucek, P.; Caraco, Y.; Guengerich, F. P.; Wood, A. J. J. *Biochem. Pharmacol.* **1996**, *53*, 111; (e) Gentile, D. M.; Tomlinson, E. S.; Maggs, J. L.; Park, B. K.; Back, D. J. *J. Pharmacol. Exp. Ther.* **1996**, *277*, 105; (f) Baarnhielm, C.; Dahlback, H.; Skanberg, I. *Acta Pharmacol. Toxicol.* **1986**, *59*, 113; (g) Lin, J. H.; Chiba, M.; Balani, S. K.; Chen, I.-W.; Kwei, G. Y.-S.; Vastag, K. J.; Nishime, J. A. *Drug Metab. Dispos.* **1996**, *24*, 1111; (h) Jang, G. R.; Wrighton, S. A.; Benet, L. Z. *Biochem. Pharmacol.* **1996**, *52*, 753; (i) Jang, G. R.; Benet, L. Z. *Drug Metab. Dispos.* **1997**, *25*, 1119; (j) Sattler, M.; Guengerich, F. P.; Yun, C.-H.; Christains, U.; Sewing, K.-F. *Drug Metab. Dispos.* **1992**, *20*, 753; (k) Fitzsimmons, M. E.; Collins, J. M. *ISSX Proc.* **1996**, *10*, 400; (l) Jacolot, F.; Simon, I.; Dreano, Y.; Beaune, P.; Riche, C.; Berthou, F. *Biochem. Pharmacol.* **1991**, *41*, 1911; (m) Surapaneni, S. S.; Clay, M. P.; Spangle, L. A.; Paschal, J. W.; Lindstrom, T. D. *Drug Metab. Dispos.* **1997**, *25*, 1383; (n) Hesse, L. M.; Venkatakrishnan, K.; von Moltke, L. L.; Shader, R. I.; Greenblatt, D. J. *Drug Metab. Dispos.* **2001**, *29*, 133; (o) Machinist, J. M.; Mayer, M. D.; Shet, M. S.; Ferrero, J. L.; Rodrigues, A. D. *Drug Metab. Dispos.* **1995**, *23*, 1163; (p) Zhao, G.; Allis, J. W. *Chem. Biol. Interact.* **2002**, *140*, 155; (q) Venkatakrishnan, K.; von Moltke, L. L.; Greenblatt, D. J. *J. Pharmacol. Exp. Ther.* **2001**, *297*, 326; (r) von Moltke, L. L.; Greenblatt, D. J.; Grassi, J. M.; Granda, B. W.; Venkatakrishnan, K.; Duan, S. X.; Fogelman, S. M.; Harmatz, J. S.; Shader, R. I. *Biol. Psychiat.* **1999**, *46*, 839; (s) Pichard, L.; Fabre, I.; Fabre, G.; Domergue, J.; Aubert, B. S.; Mourad, G.; Maurel, P. *Drug Metab. Dispos.* **1990**, *18*, 595; (t) Shimada, T.; Tsumura, F.; Yamazaki, H. *Drug Metab. Dispos.* **1999**, *27*, 1274; (u) Kalgutkar, A. S.; Taylor, T. J.; Venkatakrishnan, K.; Isin, E. M. *Drug Metab. Dispos.* **2003**, *31*, 243; (v) Zhang, Y.; Guo, X.; Lin, E. T.; Benet, L. Z. *Drug Metab. Dispos.* **1998**, *26*, 360; (w) Silvers, K. J.; Chazinski, T.; McManus, M. E.; Bauer, S. L.; Gonzalez, F. J.; Gelboin, H. V.; Maurel, P.; Howard, P. C. *Cancer Res.* **1992**, *52*, 6237; (x) Desta, Z.; Kerbusch, T.; Soukhova, N.; Richard, E.; Ko, J. W.; Flockhart, D. A. *J. Pharmacol. Exp. Ther.* **1998**, *285*, 428; (y) Kumar, G. I.; Rodrigues, A. D.; Buko, A. M.; Denissen, J. F. *J. Pharmacol. Exp. Toxicol.* **1996**, *277*, 423; (z) Prueksaritanont, T.; Gorham, L. M.; Ma, B.; Liu, L.; Yu, X.; Zhao, J. J.; Slaughter, D. E.; Arison, B. H.; Vyas, K. P. *Drug Metab. Dispos.* **1997**, *25*, 1191; (aa) Kumar, G. N.; Walle, U. K.; Walle, T. *J. Pharmacol. Exp. Ther.* **1994**, *268*, 1160; (ab) Wienkers, L. C.; Steenwyk, R. C.; Sanders, P. E.; Pearson, P. G. *J. Pharmacol. Exp. Ther.* **1996**, *277*, 982; (ac) Zhou-Pan, X.-R.; Seree, E.; Zhou, X.-J.; Placidi, M.; Maurel, P.; Barra, Y.; Rahmani, R. *Cancer Res.* **1993**, *53*, 5121; (ad) Kawashiro, T.; Yamashita, K.; Zhao, X. J.; Koyama, E.; Tani, M.; Chiba, K.; Ishizaki, T. *J. Pharmacol. Exp. Ther.* **1998**, *286*, 1294; (ae) Ohyama, K.; Nakajima, M.; Nakamura, S.; Shimada, N.; Yamazaki, H.; Yokoi, T. *Drug Metab. Dispos.* **2000**, *28*, 1303; (af) Gantenbein, M.; Attolini, L.; Bruguerolle, B.; Villard, P. H.; Puyoou, F.; Durand, A.; Lacarelle, B.; Hardwigsen, J.; Le-Treut, Y. P. *Drug Metab. Dispos.* **2000**, *28*, 383; (ag1) Eiermann, B.; Engel, G.; Johansson, I.; Zanger, U. M.; Bertilsson, L. *Br. J. Clin. Pharmacol.* **1997**, *44*, 439; (ag2) Linnet, K.; Olesen, O. V. *Drug Metab. Dispos.* **1997**, *25*, 1379; (ah) Irshaid, Y.; Adedoyn, A.; Lotze, M.; Branch, R. A. *Drug Metab. Dispos.* **1994**, *22*, 161; (ai) Obach, R. S. *Drug Metab. Dispos.* **2001**, *29*, 1057; (aj) Koyama, E.; Chiba, K.; Tani, M.; Ishizaki, T. *J. Pharmacol. Exp. Ther.* **1997**, *281*, 1199; (ak) Andersson, T.; Miners, J. O.; Veronese, M. E.; Tassaneeyakul, W.; Tassaneeyakul, W.; Meyer, U. A.; Birkett, D. J. *Br. J. Clin. Pharmacol.* **1993**, *36*, 521; (al) Zhao, X.-J.; Yokoyama, H.; Chiba, K.; Wanwimolruk, S.; Ishizaki, T. *J. Pharmacol. Exp. Ther.* **1996**, *279*, 1327; (am) Kroemer, H. K.; Echizen, H.; Heidemann, H.; Eichelbaum, M. *J. Pharmacol. Exp. Ther.* **1992**, *260*, 1052; (an) Shiraga, T.; Matsuda, H.; Nagase, K.; Iwasaki, K.; Noda, K.; Yamazaki, H.; Shimada, T.; Funae, Y. *Biochem. Pharmacol.* **1994**, *47*, 727; (ao) Marre, F.; Sanderink, G.-J.; de Sousa, G.; Gaillard, C.; Martinet, M.; Rahmani, R. *Cancer Res.* **1996**, *56*, 1296; (ap) Yamazaki, H.; Shibata, A.; Suzuki, M.; Nakajima, M.; Shimada, N.; Guengerich, F. P.; Yokoi, T. *Drug Metab. Dispos.* **1999**, *27*, 1260; (aq) Ring, B. J.; Parli, C. J.; George, M. C.; Wrighton, S. A. *Drug Metab. Dispos.* **1994**, *22*, 352.
11. Guengerich, F. P. In *Cytochrome P450: Structure, Mechanism and Biochemistry*; OrtizdeMontellano, P. R., Ed.; Plenum Press: New York, NY, 1995, pp 473–535.
12. Roussel, F.; Khan, K.; Halpert, J. R. *Arch. Biochem. Biophys.* **2000**, *374*, 269.
13. Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: London, 1999.

14. Sun, H. *Curr. Comput. Aid. Drug Des.* **2005**, *1*, 65.
15. Neal, R. *Bayesian Learning for Neural Networks*; Springer: New York, 1996.
16. Burden, F. R.; Winkler, D. A. *J. Med. Chem.* **1999**, *42*, 3183.
17. Husmeier, D.; Penny, W. D.; Roberts, S. J. *Neural Networks* **1999**, *12*, 677.
18. MacKay, D. J. C. *Neural Comput.* **1992**, *4*, 415.
19. MacKay, D. J. C. *Neural Comput.* **1992**, *4*, 448.
20. Butine, W. L.; Weigend, A. S. *Complex Syst.* **1991**, *5*, 603.
21. MacKay, D. J. C. *Comput. Neural Syst.* **1995**, *6*, 469.
22. Bishop, C. M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1995.
23. Foresee, F. D.; Hagan, M. T. IEEE, 1997, New York, pp 1930–1935.
24. (a) Wang, R. W.; Newton, D. J.; Scheri, T. D.; Lu, A. Y. H. *Drug Metab. Dispos* **1997**, *25*, 502; (b) Schmider, J.; Greenblatt, D. J.; vin Moltke, L. L.; Harmatz, J. S.; Duan, S. X.; Karsov, D.; Shader, R. I. *Pharmacology* **1996**, *52*, 125; (c) Kudo, S.; Odomi, M. *Eur. J. Clin. Pharmacol.* **1998**, *54*, 253; (d) Abel, S. M.; Back, D. J. *J. Steriod Biochem. Mol. Biol.* **1993**, *46*, 827; (e) Sutton, D.; Butler, A. M.; Nadin, L.; Murray, M. *J. Pharmacol. Exp. Ther* **1997**, *282*, 294; (f) Feierman, D. E.; Lasker, J. M. *Drug Metab. Dispos.* **1996**, *24*, 932; (g) Haaz, M.-C.; Rivory, L.; Riche, C.; Vernillet, L.; Robert *J. Cancer Res.* **1998**, *58*, 468; (h) Kuriya, S.; Ohmori, S.; Hino, M.; Ishii, I.; Nakamura, H.; Senda, C.; Igarashi, T.; Kiuchi, M.; Kitada, M. *Drug Metab. Dispos.* **2000**, *28*, 73; (i) Iatsimiskaia, E.; Tiulebaev, S.; Storozhuk, E.; Utkin, I.; Smith, D.; Gerber, N.; Koudriakova, T. *Clin. Pharmacol. Ther.* **1997**, *61*, 554; (j) Miranda, C. L.; Reed, R. I.; Guengerich, F. P.; Buhler, D. R. *Carcinogenesis* **1991**, *12*, 515; (k) Araya, Z.; Wikvall, K. *Biochim. Biophys. Acta* **1999**, *1483*, 47; (l) Rodrigues, A. D.; Mulford, D. J.; Lee, R. D.; Surber, B. W.; Kukulka, M. J.; Ferrero, J. L.; Thomas, S. B.; Shet, M. S.; Estabrook, R. W. *Drug Metab. Dispos.* **1995**, *23*, 765; (m) Fogelman, S. M.; Schimder, J.; Venkatakrishnan, K.; von Moltke, L. *Neuropsychopharmacology* **1999**, *20*, 480.